DATA PUBLICATION IN THE OPEN ACCESS INITIATIVE

Jens Klump[1]*, Roland Bertelmann[2], Jan Brase[3], Michael Diepenbroek[4], Hannes Grobe[5], Heinke Höck[6], Michael Lautenschlager[6], Uwe Schindler[4], Irina Sens[7] and Joachim Wächter[1]

[1] GeoForschungsZentrum Potsdam, Potsdam
[2] Library of the Wissenschaftspark "Albert Einstein", Potsdam
[3] Research Center L3S, University of Hannover
[4] World Data Center for Marine Environmental Sciences, (WDC-MARE), Bremen/Bremerhaven
[5] Alfred Wegener Institute for Polar and Marine Research, Bremerhaven
[6] German National Library of Science and Technology (TIB), Hannover
[7] World Data Center Climate, Max-Planck-Institut für Meteorologie, Hamburg
* Corresponding author, Email: jens.klump@gfz-potsdam.de

ABSTRACT

The 'Berlin Declaration' was published in 2003 as a guideline to policy makers to promote the Internet as a functional instrument for a global scientific knowledge base. Because knowledge is derived from data, the principles of the 'Berlin Declaration' should apply to data as well. Today, access to scientific data is hampered by structural deficits in the publication process. Data publication needs to offer authors an incentive to publish data through long-term repositories. Data publication also requires an adequate licence model that protects the intellectual property rights of the author while allowing further use of the data by the scientific community.

Keywords: *open access, access to data, science policy, intellectual property rights, data publication.*

1 DATA PUBLICATION TODAY

On 22 October 2003, a group of leading research institutions and research funding institutions published the 'Berlin Declaration on Access to Knowledge in the Sciences and Humanities' in order to "[…] promote the Internet as a functional instrument for a global scientific knowledge base and human reflection and to specify measures which research policy makers, research institutions, funding agencies, libraries, archives and museums need to consider." (Berlin Declaration, 2003). The Berlin Declaration has since been signed by 129 scientific bodies worldwide. The OECD Governments have also recognised the importance of open access to knowledge. This new policy has been formulated in the 'Communiqué on Science, Technology and Innovation for the 21st Century' issued at the OECD ministerial meeting, 29-30 January 2004 (OECD, 2004).

Knowledge, as published through scientific literature, is the last step in a process originating from primary scientific data. These data are analysed, synthesised, interpreted, and the outcome of this process is published as a scientific article. The Berlin Declaration and the OECD Ministerial Communiqué look at the outcome of this process. Because scientific knowledge is ultimately derived from data, we wish to examine more closely the beginning of this process, the issues of data sharing and data publication.

Some organisations encourage scientists to share data freely and even make data sharing a part of their funding policy (e.g. NIH, 2003). In addition, cases of scientific misconduct in recent years have highlighted the importance of making scientific data available. As a consequence, the German Science Foundation, and other science organisations, adopted 'Recommendations for Good Scientific Practice' as part of their policy. They require that institutes archive data, which were used as a basis of a publication, on safe storage media for a minimum duration of ten years (DFG, 1998). Besides being a matter of common sense and good scientific conduct, thorough documentation of experiments also makes economic sense (Alexander, Berlin, Cyr, Schofield & Platt, 2004).

The cited policy papers are concerned with archiving data as a means of safeguarding transparency; they do not mention access to data. Only a very small proportion of the original data are published in conventional scientific journals. Existing policies on data archiving notwithstanding, in today's practice data are primarily stored in private files, not in secure institutional repositories, and effectively are lost (Figure 1). This lack of access to scientific data is an obstacle to interdisciplinary and international research. It causes unnecessary duplication of research efforts, and the verification of results becomes difficult, if not impossible (Dittert, Diepenbroek &

Grobe, 2001). Large amounts of research funds are spent every year, while already existing data remain underutilised (Arzberger, Schroeder, Beaulieu, Bowker, Casey, Laaksonen et al., 2004).

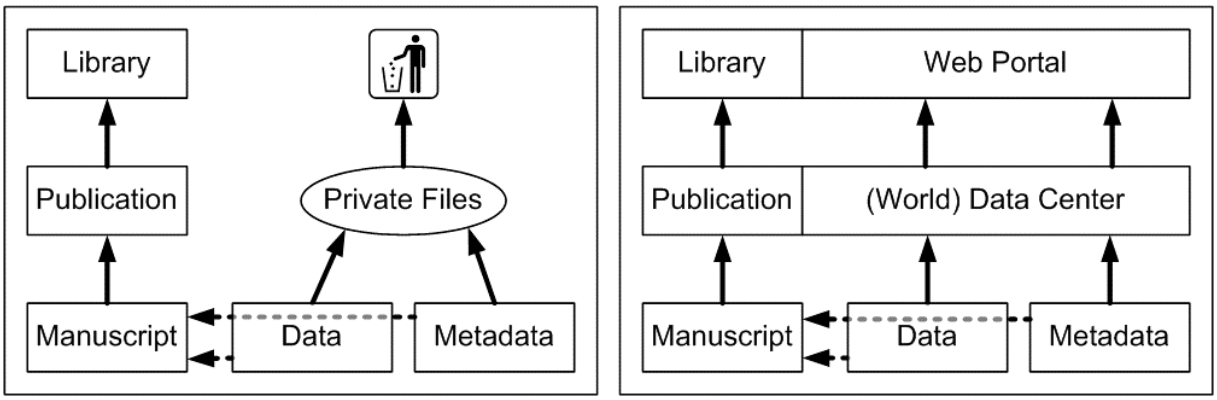

Figure 1. (left) Schematic depiction of the flow of scientific information from research to published library resources as currently practiced (modified after Helly, Staudigel & Koppers, 2003). (right) Potential approach based on data publication by data centres and content syndication to scientific web portals, which could also be library catalogues. Inter-linking publications and their underlying data will create new scientific products with added value. The dashed lines from data and metadata to the manuscript reflect the limited publication of these sources in our conventional scientific journals.

The scientific discourse today is hampered by structural problems in the publication process. The size of the data sets used in a scientific publication often prohibits their publication as printed data tables and, as a result, data used as the basis of a publication are rarely published anymore. This is surprising, since data sets are rarely too large to be transferred over the Internet. The problem at present is a lack of suitable databases and little motivation for authors to go through the work necessary to prepare their data for publication and for import into a database.

2 PRE-REQUISITES FOR DATA PUBLICATION

To make data centres and scientific web portals effective ways of data sharing, scientists need to be convinced that preparing their data for online publication is a worthwhile effort. It would be an incentive to the author if a data publication had the rank of a citeable publication, adding to his reputation and ranking among his peers. To achieve the rank of a publication, a data publication needs to meet the two main criteria, persistence and quality.

a) For data to be citeable it is necessary that they can be referred to in a persistent way. Simply making data available though the 'web' is not enough. The location of internet resources, and thus their URL, may easily change, which in most cases means to the user that the data are lost (Koehler, 2004; Lawrence, Coetzee, Glover, Pennock, Flake, Nielsen et al., 2001). Therefore, a prerequisite for data access via the internet is the use of persistent identifiers, such as DOI or URN, to have an address to locate the desired dataset that is reliable and available over a long time (Brase, 2004; Paskin, 2004).

Another aspect of persistence is that the data are stored in repositories that guarantee long-term operation. This condition is met by modern data centres, some of which are part of the ICSU system of World Data Centers, which make data accessible through their web portals (Lautenschlager, 2004).

b) Whereas persistence is mainly a technological question, data quality is a far more difficult concept. In ISO 9000:2000, the term "quality" is defined as the "degree to which a set of inherent characteristics fulfils requirements" (ISO, 2000). In terms of data, these could be credibility, usability and interpretability (Hinrichs & Aden, 2001). Defining and safeguarding technical data quality should be made part of the workflow of data integration in the data centres. To discuss scientific data quality management at length is beyond the scope of this paper.

3 OPEN ACCESS AND INTELLECTUAL PROPERTY RIGHTS

The Open Access Initiative defines the following criteria for open access:

- Irrevocable free access, worldwide,
- The licence to copy, use, distribute, transmit and display the work publicly,
- The licence to make and distribute derivative works if proper attribution of authorship is given,
- Availability through at least one online repository with long-term archiving capability.

The criteria of accessibility and long-term persistence are met by modern data centres with online access and by the use of persistent identifiers for digital data objects. In addition, publication of scientific data also requires that the intellectual property rights of the data author are guarded by an adequate licence model that allows open access to the data within the boundaries of 'fair use', including the right to produce derivative works.

Intellectual property rights and fair-use have become intensely debated issues of the "Internet Age" and the electronic distribution of data therefore needs a licensing system that supports the idea of Open Access to scientific data, yet guards the intellectual property rights of their originators. 'Fair Use' is an issue in the 'Berlin Declaration' and was discussed at the 2. Berlin Declaration Conference in May 2004 at CERN, Geneva. Here, Schlögl & Velden (2004) recommended the Creative Commons Licence System (Creative Commons, 2001) in their roadmap proposal as an appropriate licence system for publications in the sciences and humanities.

The Creative Commons Licence System is a toolbox to assemble a licence tailored to the requirements of the author. The system defines six main licensing types ranging from restrictive (requires attribution to the author, allows no commercial use or derivative works) to accommodating (only requires attribution to the author). Traditional scientific publications, in the sense of Open Access, may be downloaded and shared freely, as long as they are properly attributed. It would be considered bad scientific practice if someone produced a derivative work without proper attribution of authorship. The publication may, however, not be changed in any way and may not be used commercially. The appropriate Creative Commons licence would be by attribution, non-commercial, no derivatives (by-nc-nd) (Clarke, 2005).

Data publications should be treated in analogy to 'traditional' publications. They may be downloaded and shared freely, as long as they are properly attributed and they may not be used commercially. The difference between a scientific publication and a scientific data publication is most pronounced in the question of derivative works. It lies in the nature of a dataset that it is intended to be used in derivative works, i.e. interpretations or re-analysis of the dataset. To further the idea of Open Access, authors of a derivative work should be required to publish it under the same licence, so any derivatives will also be non-commercial in nature. Therefore, the appropriate Creative Commons licence would be by attribution, non-commercial, share alike (by-nc-sa).

Science Commons is an exploratory project, proposed in 2002 and launched in 2005, to apply the philosophy and activities of Creative Commons in the realm of science (Science Commons, 2005). Science Commons works in three project areas: Publishing, Licensing, and Data. It was and since then has differentiated into bundle of specific licences, but it is still work in progress. Science Commons is hosted at MIT's Computer Science and Artificial Intelligence Laboratory and is backed by 31 associates, among them Rice University, Harvard Law School, MIT, O'Reilly Publishers, and the Public Library of Science.

4 THE PROJECT "PUBLICATION AND CITATION OF SCIENTIFIC PRIMARY DATA"

The German CODATA group initiated a project on publication and citation of scientific data which was funded by the German Science Foundation DFG for the period 2003-2005 (STD-DOI, 2003). This project uses persistent identifiers (both DOI and URN) to identify datasets available in a digital format. The identifier is resolved to the valid location (URL) where the this dataset can be found. This approach meets one of the prerequisites for citeability of scientific data published online. In addition, the data publications are included into the catalogue of the German National Library of Science and Technology (TIB) (Brase, 2004).

In the project STD-DOI, the TIB acts as a registration agency for persistent identifiers. For every data publication, it requests a set of metadata to be incorporated into the library catalogue. The data sources are the participating World Data Centers in Germany, WDC-MARE (Bremen/Bremerhaven), WDC Climate (Hamburg) and the proposed WDC-TERRA (Potsdam). The consortium is soon to be joined by WDC-RSAT (Oberpfaffenhofen). The data centres act as registration agents for scientific and technical data DOIs. These data

centres are also responsible for technical quality control in their data domains, at the same time they also act as long-term archives. The project participants thus encompass all functions necessary for the publication of scientific data.

On May 1st 2005 the TIB became the world's first DOI registration agency for scientific primary data, working in cooperation with the World Data Center Climate (WDCC) at the Max Planck Institute for Meteorology Hamburg, GeoForschungsZentrum Potsdam, World Data Center for Marine Environmental Sciences (WDC-MARE) at the Alfred Wegener Institute Bremerhaven and at the University of Bremen and technically advised by the Research Center L3S Hannover. Through this project, the foundations have been laid for a system of scientific data publication.

5 CONCLUSIONS

Scientific knowledge is communicated through scientific literature. Knowledge is ultimately derived from data. Therefore, the 'Berlin Declaration' and the OECD Communiqué are to be applied to scientific data in the same way as they were formulated for scientific literature. Applying the 'Berlin Declaration' to data requires a publication system for data beyond 'traditional' media. The criteria of accessibility, persistent identification and long-term availability need to be met to comply with the principles of Open Access . The project 'Publication and citation of scientific primary data' (STD-DOI) shows prototypically how these criteria can be met and implements a system for the publication of scientific data, which is open to the scientific community in any scientific field.

A publication system for scientific data needs to be supplemented by an adequate licence model that allows scientists to use the published data, create new works derived from the original data, and in turn publish their new works based on these data, always respecting the intellectual property rights of the original author and the principles of 'fair use'. The options available in the Creative Commons Licence System suit many fields of scientific research. A Science Commons Licence System is desirable and necessary, especially in applied research, but modifications and alterations are still in progress.

6 REFERENCES

Alexander, W., Berlin, J., Cyr, P., Schofield, A. & Platt, L. (2004) Realities at the leading edge of research - Good practice and proper conduct in research pay off, scientifically and economically. *EMBO Reports 5* (4), 324-329. doi:10.1038/sj.embor.7400137.

Arzberger, P., Schroeder, P., Beaulieu, A., Bowker, G., Casey, K., Laaksonen, L., Moorman, D., Uhlir, P. & Wouters, P. (2004) Promoting Access to Public Research Data for Scientific, Economic, and Social Development. *Data Science Journal 3*, 135-152. http://journals.eecs.qub.ac.uk/codata/Journal/contents/3_04/3_04pdfs/DS377.pdf.

Berlin Declaration (2003) Berlin Declaration on Open Access to Knowledge in the Sciences and Humanities. Berlin: http://www.zim.mpg.de/openaccess-berlin/berlindeclaration.html.

Brase, J. (2004) Using Digital Library Techniques - Registration of Scientific Primary Data. *Lecture Notes in Computer Science 3232*, 488-494. http://www.springerlink.com/openurl.asp?genre=article&issn=0302-9743&volume=3232&spage=488.

Clarke, R. (2005) A proposal for an open content licence for research paper (Pr)ePrints. *First Monday 10* (8), 1-11. http://www.firstmonday.org/issues/issue10_8/clarke/.

Creative Commons (2001) "Some Rights Reserved": Building a Layer of Reasonable Copyright. Retrieved 2005-09-14 from the World Wide Web: http://creativecommons.org.

DFG (1998) Regeln guter wissenschaftlicher Praxis. Deutsche Forschungsgemeinschaft: http://www.dfg.de/aktuelles_presse/reden_stellungnahmen/download/self_regulation_98.pdf.

Dittert, N., Diepenbroek, M. & Grobe, H. (2001) Scientific data must be made available to all. *Nature 414* (6862), 393. doi:10.1038/35106716.

Helly, J., Staudigel, H. & Koppers, A. (2003) Scalable models of data sharing in Earth sciences. *Geochemistry, Geophysics, Geosystems - G (super 3) 4* (1), 14. doi:10.1029/2002GC000318.

Hinrichs, H. & Aden, T. (2001) An ISO 9001:2000 compliant quality management system for data integration in data warehouse systems. *International Workshop on Design and Management of Data Warehouses*, Interlaken, Switzerland.

ISO (2000) ISO 9000:2000: Quality management systems - Fundamentals and vocabulary. 2. Retrieved 2005-09-20 from the World Wide Web: http://www.iso.org/iso/en/CatalogueDetailPage.CatalogueDetail?CSNUMBER=29280.

Koehler, W. (2004) A longitudinal study of Web pages continued: a report after six years. *Information Research 9* (2). http://informationr.net/ir/9-2/paper174.html.

Lautenschlager, M. (2004) WDC Network for Earth System Research. *19th International CODATA Conference*, Berlin, Germany.

Lawrence, S., Coetzee, F., Glover, E., Pennock, D., Flake, G., Nielsen, F., Krovetz, R., Kruger, A. & Giles, L. (2001) Persistence of Web References in Scientific Research. *IEEE Computer 34* (2), 26-31. http://www.fravia.com/library/persistence-computer01.pdf.

NIH (2003) Final NIH Statement on Data Sharing. National Institute of Health, Bethesda, MD: http://grants2.nih.gov/grants/guide/notice-files/NOT-OD-03-032.html.

OECD (2004) Science, Technology and Innovation for the 21st Century. Meeting of the OECD Committee for Scientific and Technological Policy at Ministerial Level, 29-30 January 2004 - Final Communique. Organisation for Economic Co-operation and Development, Paris, France: http://www.oecd.org/document/0,2340,en_2649_34487_25998799_1_1_1_1,00.html.

Paskin, N. (2004) Digital Object Identifiers for scientific data sets. *19th International CODATA Conference*, Berlin, Germany.

Schlögl, R. & Velden, T. (2004) Berlin 2 Open Access - Roadmap Proposal. *Berlin 2 Open Access: Steps Toward Implementation of the Berlin Declaration on Open Access to Knowledge in the Sciences and Humanities*, CERN, Geneva.

Science Commons (2005) Science Commons. Retrieved 2005-09-20 from the World Wide Web: http://sciencecommons.org.

STD-DOI (2003) Publication and Citation of Scientific Primary Data. Retrieved 2005-09-20 from the World Wide Web: http://www.std-doi.de.